图书情报工作
第 64 卷 第 14 期 2020 年 7 月

融合论文关键词知识的专利术语抽取方法*

■ 俞琰[1,2] 陈磊[1] 姜金德[3] 赵乃瑄[1]

[1] 南京工业大学信息服务部 南京 210009 [2] 东南大学成贤学院计算机工程系 南京 211816

[3] 南京晓庄学院商学院 南京 211171

摘 要：[目的/意义]提出利用丰富的论文关键词知识获取专利文本之外的有效特征，以弥补因专利文本集自身信息不足而制约专利术语抽取效果这一缺陷，提高专利术语抽取准确率。[方法/过程]根据相关论文的关键词知识，分别提出领域相关度和首尾度两个特征，以衡量候选术语成为术语的可能性，并将这些特征融入到专利术语抽取的传统方法之中。[结果/结论]实验结果表明，利用论文关键词得到的候选术语领域相关度和首尾度信息，可使结合论文关键词知识的方法比传统的术语抽取方法的准确率有了明显的提升。

关键词：专利术语抽取 论文 关键词

分类号：G202

DOI：10.13266/j.issn.0252－3116.2020.14.011

1 引言

专利文献是技术信息的重要来源，有效的专利文献分析对国家经济、科技、民生的发展起着至关重要的作用。其中，专利文献中的术语为专利文献分析提供了结构化知识单元，体现和承载了专利文献的技术信息，成为诸多专利文献分析的关键组成部分。因此，从专利文献中通过无人工干预或尽量少的人工干预的方法自动抽取专利术语是一个重要的研究课题。

C-value 方法[1]是一种常用的基于统计的术语抽取方法，在长术语抽取方面表现较好。然而，C-value 方法主要基于术语频次计算，存在低频术语无法被识别（如，词串“功能化 石墨 烯”因在专利文本集中出现次数较少，而没有被正确抽取）以及部分边界识别不正确（如，包含边界词“通入”的词串“通入 惰性 气体”因在专利文本集中出现次数较多，而被错误地抽取）等问题[2]，抽取准确率仍有较大提升空间。

高质量论文是科学研究的主要输出形式，是专利的主要理论来源与知识源泉[3]。相应地，专利是技术创新的成果体现，为科学研究启示问题、拓展研究空间、激发创新灵感。特别是近年来，科学研究和技术创新之间的交互作用日益活跃，两者之间的关系愈发紧密，使得论文和专利具有较强的相关性。论文中通常包含作者标引的描述全文主题内容的关键词。关键词标引不是随意的，一般为特定领域成熟术语或词组[4－5]。因此，为了弥补因专利语料自身的信息不足而制约专利术语抽取效果这一缺陷，本文首次提出利用丰富的论文关键词知识获取专利文本之外的有效特征，以提高专利术语抽取效果。方法是根据相关论文的关键词知识，分别提出两类特征衡量候选术语成为术语的可能性，并将这些特征融入到 C-value 方法之中，以提高专利术语抽取的准确率。

2 相关工作

2.1 术语抽取

目前的术语抽取方法可分为基于统计的方法和基于机器学习的方法两大类。

基于统计的方法通过计算统计量来评估词串成为术语的可能性，具有较少人工干预、较强的适应性和可移植性等优点，一般使用术语性和单元性度量候选术语成为术语的可能性。术语性从术语的隶属度出发，衡量一个候选术语与特定领域的相关程度。常用的术

* 本文系国家社会科学基金一般规划项目“大数据时代支持创新设计的多维度多层次专利文本挖掘研究”（项目编号：17BTQ059）研究成果之一。

作者简介：俞琰（ORCID：0000－0002－9654－8614），教授，博士，E-mail：yuyanyuyan2004@126.com；陈磊，硕士研究生；姜金德（ORCID：0000－0002－5504－7493），教授，博士；赵乃瑄（ORCID：0000－0001－9072－7315），教授，博士。

收稿日期：2018－12－23 修回日期：2019－10－05 本文起止页码：104－111 本文责任编辑：杜杏叶

语性统计量有词频[6]、TF-IDF[7]、C-value[1]方法等。该类方法主要基于术语频次计算,存在低频词无法被识别以及部分边界识别不正确等问提。针对这些问题,目前也有一些改进研究[8－13],总的来说,较为典型的改进方法包括引入互信息和邻接熵[14]两种统计量,以重构目标函数。然而,研究结果表明,这些方法的改进仍有较大的提升空间。单元性度量候选术语结构的稳定程度,即候选术语内部各组成部分之间的结合强度。其中,互信息是一种常用的单元性指标[15],互信息通过计算候选术语中各词成分的共现频次来衡量这些成分之间的依赖程度,能够较好地反映字串之间的结合强度,但会过高估计低频且总是相邻出现的字串间的强度。一些研究尝试改进这个问题[16－17],但结果仍有较大提升空间。

基于机器学习的方法通过学习训练文本特征构造模型抽取术语。基于机器学习的方法可以弥补基于统计的方法无法识别低频术语的缺陷,利用数据学习模型,判断词串成为术语的可能性。常用的机器学习方法包括最大熵模型[18]、条件随机场模型[19－20]等。但基于机器学习的方法需要大规模人工标注语料作为训练数据,对训练语料的规模与质量要求较高,并且,基于机器学习的方法还不成熟,需要进行更多的尝试与验证[21]。目前专利文献没有有针对性的、完备的、大规模标注语料,基于统计的方法可以在极低人工干预下抽取术语,是克服基于机器学习的方法中标注语料获取困难的有效途径之一。因此,本文着重研究使用统计的方法抽取专利术语。

2.2 论文与专利的相关性

科学研究与技术发明相互作用,在知识传递与反馈中旋进发展[22]。近年来,国内外相关研究表明论文和专利具有较强的相关性。

在国外,F. Narin 等[23]选取生物医学杂志以及美国专利商标局专利数据库中与生物技术相关的专利,分析论文与论文的引用关系、专利与专利的引用关系、以及专利与论文间的引用关系,揭示高科技技术与科学之间的关系十分紧密,论文与专利间具有较强的相关性。F. Narin 等还发现科学与技术建的知识关联程度每 6 年会增加 2 倍[24]。T. Magerman 等[25]以专利发明人发明者身份与论文作者双重身份,或专利发明人和科研学者共同合作关系为切入点,使用 LSA 文本挖掘方法发现专利文献与论文间存在较高的相似性。Y. Qi 等[26]通过大规模收集纳米科学领域专利和论文,利用主题关键词提取语义级主题,揭示了论文与专利之间的相关性。H. Huang 等[27]分析了燃料电池领域论文和专利的交叉引用情况,表明燃料电池领域的科学与技术关联呈现逐渐增加的收敛性。

在国内,吴菲菲等[28]通过论文与专利之间的引用关系,结合社会网络分析方法,发现科学领域和技术领域之间存在相互作用关系。特别地,近十年来化学、通讯、计算机、医疗器械、测量等领域技术对科学影响很大;化学、物理、生物、医学等领域科学研究成果对专利成果的形成具有普遍影响。彭彦淇等[29]使用引文分析法与专利计量法对石墨烯领域专利和论文进行交叉引用分析,揭示了该领域中科学与技术的关联性。黄鲁成等[30]运用文本挖掘方法并在完善 SAO 结构基础上,发现钙钛矿太阳能电池领域中论文和专利的相似性。

3 融合论文关键词知识的专利术语抽取方法

针对目前术语抽取存在的问题,本文提出融合论文关键词知识的专利术语抽取方法。方法流程见图 1,主要包括预处理(见 3.1 小节)、候选术语选取(见 3.2 小节)、C-value 值计算(见 3.3 小节)、基于关键词特征统计(见 3.4 小节)和 C-value 值更新(见 3.5 小节)等 5 个主要步骤。

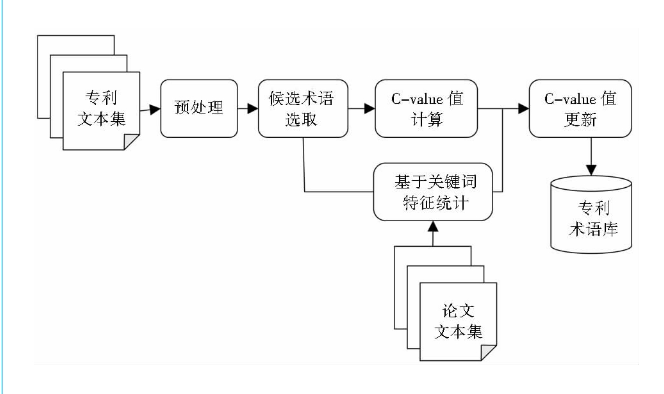


图 1　融合论文关键词知识的专利术语抽取方法流程

3.1 预处理

首先对收集的专利文本集和论文文本集进行预处理。预处理主要包括分词、词性标注、去除停用词等工作。其中,由于中文文本词与词之间没有明显的切分标记,需要通过分词把一个句子按照其中词的含义进行切分。词性标注的任务是分词后为句子中的每个词赋予一个合适的词性。去除停用词则通过通用停用词表以及人工筛选去除频率高但是信息量少的词,如

“的”“了”“发明”等词。此外,预处理工作还包括英文大小写格式转换、去除特殊符号等工作。

3.2　候选术语选取

术语中一般不包含连词、介词、助动词、副词和标点符号,因此,在候选术语选取步骤中,使用人工制定的语法规则,从语料中选取候选术语。词性模式匹配方法根据特定的词性排列模式,以选取名词短语作为候选术语,本文使用文献[9]的词性模式匹配规则选取候选术语。词性模式匹配规则如表 1 所示,其中 a 表示形容词,b 表示区别词,c 表示连词,d 表示副词,k 表示后接成分,l 表示习语,m 表示数词,n 表示名词,u 表示助词,v 表示动词,vn 表示对应位置即可是动词也可是名词,加号表示多词术语由相应词性的词组合而成。

表 1　中文专利候选术语词性模式匹配规则[9]

长度	词性模式匹配规则
2 词	n + n、n + v、v + n、a + n、d + n、b + n
3 词	n + n + n、v + n + n、n + v + n、v + v + n、b + v + n、n + m + n
4 词	n + n + n + n、n + n + v + n、v + n + n + n、v + n + v + n、n + v + v + n、v + v + n + n、v + n + b + n
5 词	v + v + n + n + n、d + v + n + n + n、m + v + m + n + n、b + v + n + v + n、n + n + v + n + n、a + n + v + n + n
6 词	n + n + c + vn + n + n、n + n + vn + c + vn + n、n + n + u + b + vn + n、vn + n + vn + c + vn + n、l + vn + k + n + vn + n、n + vn + u + n + vn + n

3.3　C-value 值计算

C-value 方法为每个候选术语计算术语性,C-value 与该候选术语在语料中的频次有关,频次越高,其术语度越大。在此基础上,又考虑了候选术语的长度,认为长串出现频次比短串出现频次更有意义,是术语的可能性更大。C-value 值计算公式如下:

$$\text{C-value}(x)=\begin{cases}\log|x|\cdot \mathrm{f}(x) & x\text{ 未被嵌套}\\ \log|x|\cdot\left(\mathrm{f}(x)-\dfrac{1}{|\mathrm{T}_x|}\sum_{y\in \mathrm{T}_x}\mathrm{f}(y)\right) & \text{其他}\end{cases}$$

公式(1)

其中,x 表示候选术语;$|x|$ 表示 x 的长度;$\mathrm{f}(x)$ 表示 x 在专利文本集中出现的频次;T_x 表示专利文本集包含 x 的候选术语集合;$|\mathrm{T}_x|$ 表示专利文本集包含 T_x 中元素个数。

3.4　基于关键词特征统计

如引言中所述,C-value 方法主要考虑候选术语出现频次这一因素,从而产生低频术语无法被识别以及部分边界识别不正确等问题。而论文和专利具有较强的相关性,且论文中关键词一般为特定领域成熟术语或词组。因此,针对 C-value 存在的两个问题,本文尝试利用论文关键词知识,分别提出候选术语的领域相关度(见 3.4.1 节)和首尾度(见 3.4.2 节)两个统计特征,以弥补 C-value 主要考虑词频因素的不足,从而提高术语的抽取准确率。

3.4.1　领域相关度

针对 C-value 无法识别低频术语的问题,本文提出利用候选术语在论文文本集中作为关键词出现的频次,衡量该候选术语的领域相关度。例如,虽然候选术语“功能化 石墨 烯”在专利文本集中出现频次较低,但其在论文文本集中作为关键词频繁出现,则表明该候选术语具有较高的领域相关度,依然可以推论该候选术语可能是术语,从而缓解 C-value 方法无法识别低频术语的问题,提高术语抽取的准确率。因此,给定候选术语 x,其领域相关度 $\mathrm{D}(x)$ 为:

$$\mathrm{D}(x)=\mathrm{N}(x)$$

公式(2)

其中,$\mathrm{N}(x)$ 表示 x 在论文文本集中作为关键词出现的频次。

然而,由于术语表达的灵活性,特别是专利申请人为了扩大所申请专利的保护范围和提高专利授权的可能性,往往使用一些模糊的术语和表达,造成论文文本集中与候选术语精确匹配的关键词有限。例如,当候选术语“化学 气相 沉积”在论文文本集中没有作为关键词出现时,则该候选术语的领域相关度为 0。因此,本文将候选术语与论文关键词的精确匹配放宽为模糊匹配,即利用与候选术语词面相似的关键词衡量该候选术语的领域相关度。如,候选术语“化学 气相 沉积”虽然没有与其精确匹配的关键词,但是可以利用“化学 气相 沉积 法”、“常压 化学 气相 沉积”等模糊匹配的相似关键词计算其领域相关度。因此,给定候选术语 x 和关键词 k,更新 x 领域相关度 $\mathrm{D}(x)$:

$$\mathrm{D}(x)=\sum_k \mathrm{sim}(x,k)\times \mathrm{N}(k)$$

公式(3)

其中,$\mathrm{N}(k)$ 表示关键词 k 在论文文本集中出现的频次,$\mathrm{sim}(x,k)$ 表示候选术语 x 与关键词 k 的相似度,使用经典的 Dice 系数衡量候选术语 x 与分词后关键词 k 间的相似度,其计算公式[31]为:

$$\mathrm{sim}(x,k)=2\times\frac{|x\cap k|}{|x|+|k|}$$

公式(4)

其中,$|x\cap k|$ 表示候选术语 x 与分词后关键词 k 中相同词语数,$|x|$ 表示候选术语 x 包含词语数,$|k|$ 表示分词后关键词 k 中词语数。例如,候选术语“化学 气相 沉积”与关键词“化学 气相 沉积 法”的相似度 $sim=2\times\frac{3}{3+4}=0.86$。表 2 为候选术语“功能化 石墨 烯”和“化学 气相 沉积”的领域相关度计算示例。由

表2可见，候选术语利用论文文本集中相似关键词出现的频次，计算其领域相关度，从而缓解低频候选术语无法被正确识别的问题。为了避免不相似的关键词干扰，本文仅考虑大于相似度阈值 δ 的关键词计算候选术语领域相关度。

表2　候选术语领域相关度计算示例

候选术语	相似关键词	相似度	关键词频次	领域相关度
功能化 石墨 烯	功能化 石墨 烯	1.00	30	1.00×30+1.00×20+0.86×10=58.6
	石墨 烯 功能化	1.00	20	
	硅 功能化 石墨 烯	0.86	10	
化学 气相 沉积	化学 气相 沉积 法	0.86	20	0.86×20+0.86×10=25.8
	常压 化学 气相 沉积	0.86	10	

计算候选术语领域相关度的伪代码如下：

算法：计算候选术语的领域相关度

输入：候选术语 x、论文关键词集K、论文集Docs

输出：候选术语 x 的领域相关度 $D(x)$

1. $D(x)=0$ //将候选术语 x 的领域相关度初始值设置为0。
2. **FOR** k **IN** K **DO** //对论文关键词集K中的每个关键词 k 做如下操作：
3. $sim(x,k)=2\times\frac{|x\cap k|}{|x|+|k|}$ //根据公式(4)计算候选术语 x 与论文关键词 k 的相似度。
4. **IF** $sim(x,k)\geqslant\delta$ **DO** //判断候选术语 x 与论文关键词 k 的相似度是否大于阈值 δ，
5. $N(k)=COUNT(k, Docs)$ //统计关键词 k 在论文集Docs中出现的次数。
6. $D(x)=D(x)+sim(x,k)\times N(k)$ //如果候选术语 x 与论文关键词 k 相似，则利用公式(3)累计求和计算候选术语 x 的领域相关度 $D(x)$。
7. **END IF**
8. **END FOR**

3.4.2　首尾度

C-value方法的第二个主要问题为部分边界术语识别不正确。例如，候选术语"通入 惰性 气体"因其在专利集中频繁出现，而被错误地将边界词"通入"作为术语的一部分。而利用论文关键词信息，可以发现"通入"一词较少作为关键词的第一个词(即，首词)，从而推论该候选术语可能具有错误的首词，为正确术语的可能性较小；类似地，利用关键词对候选术语的最后一个词(即，尾词)进行统计，推论尾词的正确性，从而估计其成为术语的可能性。因此，本文利用关键词信息，提出候选术语的首度、尾度和首尾度统计特征，评估候选术语首词和尾词的正确性，以缓解C-value部分边界术语识别不正确问题，从而提高术语抽取的准确率。

具体地，给定候选术语 $x=\{w_1, w_2,\cdots,w_n\}$，候选术语首度H、尾度T和首尾度HT定义分别如下：

$$H(x)=H(w_1,*)\qquad 公式(5)$$

$$T(x)=N(*,w_n)\qquad 公式(6)$$

$$HT(x)=\min(H(x),T(x))\qquad 公式(7)$$

其中，$N(w_1,*)$ 表示以词 w_1 作为首词的关键词频次，$N(*,w_n)$ 表示以词 w_n 作为尾词的关键词频次，$\min(H(x), T(x))$ 表示从首度 $H(x)$ 和尾度 $T(x)$ 中取较小的值，表明只要候选术语首词或者尾词可能不正确，则该候选术语就可能不是术语。表3为候选术语"荧光 纳米 颗粒"和"筛选 药物"的首尾度计算示例。由表3可见，候选专利术语"荧光 纳米 颗粒"因首词"荧光"和尾词"颗粒"均频繁出现在关键词首部和尾部，因此具有较高的首尾度，其成为术语的可能性较大；而"通入 惰性 气体"由于首词"通入"没有出现在关键词词首，使得其首尾度为最小值0，表明其成为术语的可能性较小。

表3　候选术语首尾度计算示例

候选专利术语	首词或尾词	相应关键词(频次)	H或T	HT
荧光 纳米 颗粒	首词：荧光	**荧光** 探针(40) **荧光** 适体 传感器(20)	60	40
	尾词：颗粒	纳米 **颗粒**(30) 磁性 **颗粒**(10)	40	
通入 惰性 气体	首词：通入	／(0)	0	0
	尾词：气体	挥发性 有机 **气体**(1) 可燃 **气体**(2)	3	

计算候选术语首尾度的伪代码如下：

算法：计算候选术语的首尾度

输入：候选术语 x、论文关键词集K、论文集Docs

输出：候选术语 x 的首尾度 $HT(x)$

1. $w_1, w_n=CUT(x)$ //使用分词工具切分候选术语 x，将切分后的首词设为 w_1，尾词设为 w_n。
2. $H(x)=0$ //设置候选术语 x 的首度初始值为0。
3. $T(x)=0$ //设置候选术语 x 的尾度初始值为0。
4. $HT(x)=0$ //设置候选术语 x 的首尾度初始值为0。
5. **FOR** k **IN** K **DO** //对论文关键词集K中的每个关键词 k 做如下操作
6. $k_1, k_m=CUT(k)$ //使用分词工具切分关键词 k，将切分后的首词设为 k_1，尾词设为 k_m。
7. **IF** $w_1==k_1$ **DO** //如果候选术语 x 的首词 w_1 与关键词 k 的首词 k_1 相同，则累加求和首度H。
8. $N(k)=COUNT(k, Docs)$ //统计关键词 k 在论文集Docs中出现的次数。
9. $H(x)=H(x)+N(k)$ //累加求和首度H。
10. **END IF**
11. **IF** $w_n==k_m$ **DO** //如果候选术语 x 的尾词 w_n 与关键词

k 的尾词 k_m 相同,则累加求和尾度 T。

```
N(k) = COUNT(k, Docs) //统计关键词 k 在论文集 Docs 中出现的次数。
T(x) = T(x) + N(k) //累加求和尾度 T。
  END IF
END FOR
HT(x) = min(H(x), T(x)) //根据公式(7),选择首度和尾度中较小的值作为首尾度值。
```

3.5 C-value 值更新

将基于关键词统计特征的信息融入到 C-value 之中,以提高专利术语抽取准确率。具体地,结合领域相关度 D,形成 D-C-value,其定义为:

$$\text{D-C-value}(x) = (1 + D(x)) \times \text{C-value}(x) \quad \text{公式(8)}$$

由定义可知,当候选术语 C-value 值越大且领域相关度 D 越大时,则该候选术语越可能是术语,从而缓解 C-value 方法中低频术语无法被抽取的问题。特别地,当候选术语的领域相关度为 0 时,则 D-C-value 退化为 C-value。

结合首尾度 HT,形成 HT-C-value,其定义为:

$$\text{HT-C-value}(x) = (1 + HT(x)) \times \text{C-value}(x) \quad \text{公式(9)}$$

由定义可知,当候选术语 C-value 值越大,且其首尾度 HT 越大时,则该候选术语越可能是术语,从而缓解 C-value 方法中部分边界识别不正确的问题。特别地,当候选术语的首尾度为 0 时,则 HT-C-value 退化为 C-value 方法。

同时考虑领域相关度和首尾度信息,则形成 D-HT-C-value:

$$\text{D-HT-C-value}(x) = (1 + D(x)) \times (1 + HT(x)) \times \text{C-value}(x) \quad \text{公式 (10)}$$

4 实验

4.1 数据集

为了验证提出模型的可行性与有效性,本文选取石墨烯专利文献进行实验。石墨烯是已知材料中最薄的一种,因其具有独特的结构,集优异的光学、化学、电学、力学等特征于一身,被认定为新型潜力材料,具有可观的经济效益和广泛的产业化应用前景。近几年来,石墨烯研究的论文数量与专利申请量皆呈指数增长趋势。

实验专利数据基于中国国家知识产权局专利数据库,以“石墨烯”关键词检索中国近 5 年来(2014 – 2018 年)的有效中国发明公开专利(检索日期为 2018 年 11 月 15 日),共获得 6 445 条有效中国发明公开专利,以其题名和摘要作为专利文本集。实验论文数据基于万方数据库,以“石墨烯”关键词检索近 5 年来(2014 – 2018 年)北大核心期刊论文(检索日期为 2018 年 11 月 15 日),共获得 5 236 条论文数据,获取论文关键词形成论文文本集。

4.2 评估指标

鉴于专利文本数据较多,采用准确率 P@ N 作为方法的评估指标[26]:

$$P@N = \frac{r}{N} \times 100\% \quad \text{公式(11)}$$

其中,N 为常数,表示方法抽取的专利术语数,本实验分别取 200 – 2 000;r 表示 N 个专利术语中正确的个数。为了避免主观性和领域知识的局限性,利用百度百科、维基、互动百科等知识网站查找是否存在对应的词条,以判断被抽取术语是否正确。

4.3 结果

4.3.1 领域相关度对专利术语抽取准确率的影响

实验首先研究领域相关度对专利术语抽取准确率的影响。为此,将相似度阈值 δ 分别设为 0.2、0.4、0.6、0.8 和 1.0,与传统的 C-value 方法进行比较,对应的方法分别记为 D-C-value-0.2、D-C-value-0.4、D-C-value-0.6、D-C-value-0.8 和 D-C-value-1.0。实验结果如图 2 所示:

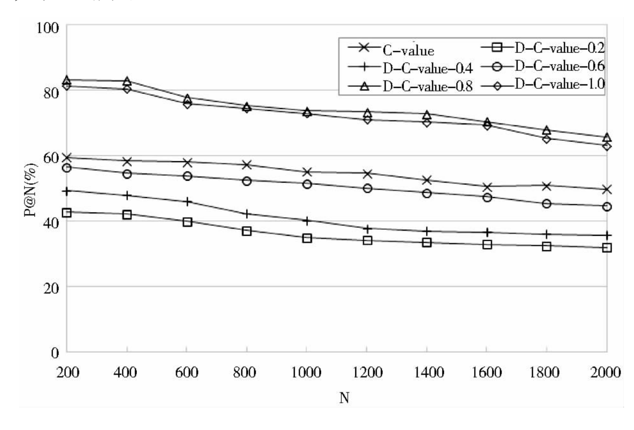


图 2 领域相关度对专利术语抽取准确率的影响

由图 2 可见,D-C-value-0.2、D-C-value-0.4 和 D-C-value-0.6 准确率低于 C-value 方法,如当 N = 1000 时,D-C-value-0.2、D-C-value-0.4 和 D-C-value-0.6 分别比 C-value 方法下降了 20.09%、14.90% 和 3.61%;而 D-C-value-0.8 和 D-C-value-1.0 准确率明显高于 C-value 方法,特别地,D-C-value-0.8 取得了最高准确率,如当 N = 1000 时,D-C-value-0.8 和 D-C-value-1.0 比 C-value 准确率提高了 18.69% 和 17.79%。这表明进行专利术语抽取时,选择与候选专利术语相似度不大的关键

词进行领域相关度特征统计,会产生噪声数据,造成准确率降低;而通过选择与候选专利术语相似度较大的关键词进行领域相关度特征统计则能够提高专利术语抽取的准确率,这也表明了利用关键词计算候选专利术语领域相关度特征对专利术语抽取的有效性,后续的实验将相似度阈值设为0.8。

4.3.3 首尾度对专利术语抽取准确率的影响

接着,实验评估首尾度对专利术语抽取的影响。为此比较传统的 C-value 术语抽取方法与添加首尾度的 HT-C-value 方法的专利术语抽取准确率。实验结果见图3。由图3 可见,包含首尾度信息的 HT-C-value 的准确率高于 C-value 方法,如当 N =1 000 时,HT-C-value 的准确率比 C-value 提高了 14.15% 。这表明通过关键词统计专利候选术语的首尾度特征,能够缓解 C-value 方法部分边界识别不正确的问题,从而提高专利术语抽取的准确率。

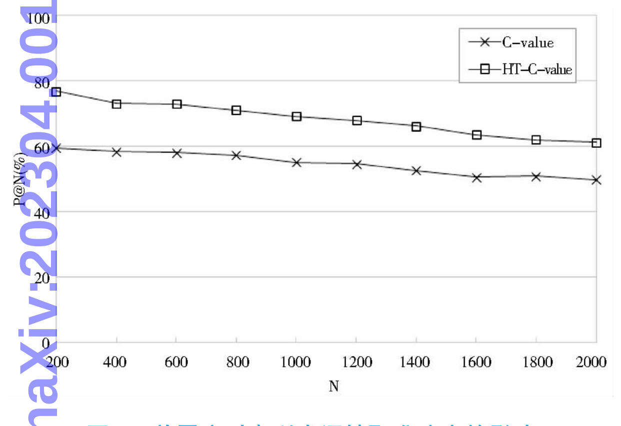


图 3　首尾度对专利术语抽取准确率的影响

4.3.4 合成特征对专利术语抽取的影响

基于上述实验,研究合成领域相关度和首尾度特征对专利术语抽取的影响。为此,将领域相关度和首尾度融入到 C-value 之中,得到 D-HT-C-value 值,与单一特征 D-C-value 和 HT-C-value 进行比较,并将 C-value 作为基准方法。实验结果见图 4。由图 4 可见,D-HT-C-value 方法准确率最高。例如,当 N =1 000 时,D-HT-C-value 的准确率比 C-value、D-C-value 和 HT-C-value 方法分别高 27.49% 、8.80% 和 13.34% 。这表明比起单一特征,将领域相关度和首尾度同时融入到 C-value 中,能够获得比单一特征融入 C-value 之中更好的准确率,特别地,在这两个特征中,领域相关度对专利术语抽取的准确率影响更大。

4.3.4 与其他方法比较

最后,使用 D-HT-C-value 方法与一些典型的 C-value 改进方法进行比较。欲比较的方法有如下几种:

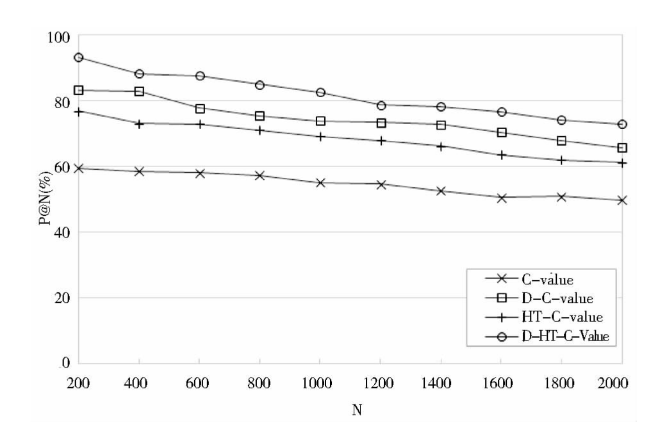


图 4　合成特征对专利术语抽取准确率的影响

(1)C-value:使用 C-value 度量候选术语术语度;

(2)PMI-C-value:将候选术语的互信息融入 C-value 之中,形成 PMI-C-value 方法。互信息是一种常用的单元性指标,通过计算候选术语中各词成分的共现频次来衡量这些成分之间的结合强度;

(3)En-C-value:将候选术语的邻接熵融入 C-value 值之中,形成 En-C-value 方法。邻接熵根据候选术语左右邻接词的不确定性消除部分候选术语边界不正确的问题。邻接熵越大,表明候选术语其邻接词包含的信息越多,其成为术语的概率越大;

(4)D-HT-C-value:本文提出的基于论文关键词,将领域相关度和首尾度融入 C-value 方法之中。

图 5 为实验结果:

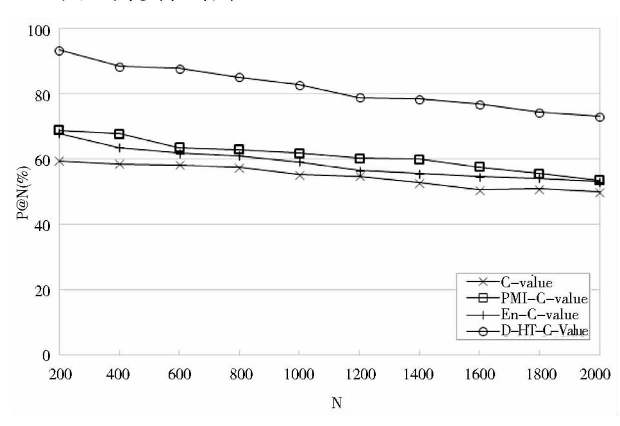


图 5　与其他方法比较专利术语抽取的准确率

由图 5 可见,PMI-C-value、En-C-value 和 D-HT-C-value 方法的准确率较 C-value 方法更高,例如当 N =1 000时,PMI-C-value、En-C-value 和 D-HT-C-value 方法的准确率比 C-value 方法提高了 6.58% 、3.89% 和 26.68% 。特别地,D-HT-C-value 方法取得了最高准确率。虽然一些低频专利候选术语可以通过互信息提高其成为候选术语的可能性,但是互信息也造成一些高

频正确术语具有较低的互信息值,从而使得术语抽取的准确率提升效果有限;而在专利术语中一些非术语高频词串反复出现,具有较多的邻接词,所以使得结合邻接熵方法抽取术语的准确率改善有限;而通过论文文本集获得候选术语的领域相关度和首尾度能较为有效地改善 C-value 的准确度,表明基于论文关键词,将领域相关度和首尾度融入 C-value 方法的有效性。

5 总结

目前的专利术语抽取方法主要存在低频术语无法被识别以及部分边界识别不正确等问题,专利术语抽取结果仍有较大的提升空间。以往的研究主要使用专利文本本身的一些特征信息,以提高专利术语抽取准确率。论文和专利具有较强的相关性,论文中关键词标引不是随意的,一般为特定领域成熟术语或词组。关键词包含丰富的特定领域知识。因此,针对目前专利术语自动抽取方法对外部资源的利用率较低的问题,为了弥补因专利文本集自身的信息不足而制约专利术语抽取效果这一缺陷,本文首次提出利用丰富的论文关键词知识获取专利文本之外的有效特征,以提高专利术语抽取效果。该方法根据相关论文的关键词知识,分别提出领域相关度和首尾度等两类特征衡量候选术语成为术语的可能性,并将这些特征融入到专利术语抽取的 C-value 方法之中,形成结合论文关键词的 C-value 方法。实验结果表明,与传统的术语抽取方法相比,结合论文关键词的方法能够有效地提高专利术语抽取的准确率。

未来的研究将尝试获取百度百科、维基、互动百科等知识网站的词条知识,以进一步提高专利文本术语抽取的准确率。

参考文献:

[1] FRANTZI K, ANANIADOU S, MIMA H. Automatic recognition of multi-word terms: the C-value/NC-value method[J]. International journal on digital libraries, 2000, 3(2): 115 – 130.
[2] 周霜霜,徐金安,陈钰枫等. 融合规则与统计的微博新词发现方法[J]. 计算机应用, 2017, 37(4):1044 – 1050.
[3] HIROYUKI T, TAKAKAYUKI T. A bibliometric analysis of scientific literatures cited by influential patents[J]. Journal of information processing and management, 2006, 49(1): 2 – 10.
[4] 陈红娟. 科技论文关键词选取[J]. 西安石油大学学报(自然科学版), 2011,26(4): 109 – 110.
[5] 李娜, 戎文慧, 边志英. 如何确定关键词[J]. 临床荟萃, 2003, 18(12):674 – 674.
[6] 覃佳慧, 何耶奇, 叶鹰. 科学论文和技术专利的引用时滞及循环周期研究[J]. 情报理论与实践, 2018, 41(7): 23 – 25.
[7] 曾文, 徐硕, 张运良, 等. 科技文献术语的自动抽取技术研究与分析[J]. 现代图书情报技术, 2014(1): 51 – 55.
[8] SPASIC I, GREENWOOD M, PREECE A, et al. FlexiTerm: a flexible term recognition method[J]. Journal of biomedical semantics, 2013, 27(4): 1 – 15.
[9] 韩红旗, 朱东华, 汪雪锋. 专利技术术语的抽取方法[J]. 情报学报, 2011, 30(12):1280 – 1285.
[10] 胡阿沛, 张静, 刘俊丽. 基于改进 C-value 方法的中文术语抽取[J]. 现代图书情报技术, 2013, 230(2): 24 – 29.
[11] 张雷瀚, 吕学强, 李卓,等. 领域本体术语的抽取方法研究[J]. 情报学报, 2014, 33(2): 167 – 174.
[12] 周霜霜, 徐金安, 陈钰枫,等. 融合规则与统计的微博新词发现方法[J]. 计算机应用, 2017, 37(4): 1044 – 1050.
[13] 俞琰, 赵乃瑄. 基于通用词与术语部件的专利术语抽取[J]. 情报学报, 2018, 37(7): 742 – 752.
[14] 丁杰, 吕学强, 刘克会. 基于边界标记集的专利文献术语抽取方法[J]. 计算机工程与科学, 2015, 37(8):1591 – 1598.
[15] 刘剑, 唐慧丰, 刘伍颖. 一种基于统计技术的中文术语抽取方法[J]. 中国科技术语, 2014, 16(5):10 – 14.
[16] 杜丽萍, 李晓戈, 于根,等. 基于互信息改进算法的新词发现对中文分词系统改进[J]. 北京大学学报(自然科学版), 2016, 52(1):35 – 40.
[17] ZHANG W, YOSHIDA T, TANG X, et al. Improving effectiveness of mutual information for substantival multiword expression extraction[J]. Expert systems with applications an international journal, 2009, 36(8):10919 – 10930.
[18] 木合亚提·尼亚孜别克, 古力沙吾利·塔里甫. 哈萨克语 IT 领域术语识别研究与实现[J]. 中文信息学报, 2016(3):68 – 73.
[19] 王昊, 王密平, 苏新宁. 面向本体学习的中文专利术语抽取研究[J]. 情报学报, 2016, 35(6):573 – 585.
[20] ZENG D, SUN C, LIN L, et al. LSTM-CRF for drug-named entity recognition[J]. Entropy, 2017, 19(6):283 – 295.
[21] CONRADO M, PARDO T, REZENDE S. A machine learning approach to automatic term extraction using a rich feature set[C]// The 2013 conference of the north American chapter of the association for computational Linguistics: human language technologies. Atlanta, Geoogia: Association for Computational Linguistics, 2013: 16 – 23.
[22] BHATTACHARYA S, KRETSCHMER H, MEYER M. Characterizing intellectual spaces between science and technology[J]. Scientometrics, 2003, 58(2): 369 – 390.
[23] NARIN F, NOMA E. Is technology becoming science? [J]. Scientometrics, 1985, 7(3): 369 – 381.
[24] NARIN F, HAMILTON K S, OLIVASTRO D. The increasing linkage between U. S. technology and public science [J]. Research policy, 1997, 26(3): 317 – 330.
[25] MAGERMAN T, LOOY B V, SONG X. Exploring the feasibility

and accuracy of latent semantic analysis based text mining techniques to detect similarity between patent documents and scientific publications[J]. Scientometrics, 2010, 82(2):289－306.
[26] QI Y, ZHU N, ZHAI Y, et al. The mutually beneficial relationship of patents and scientific literature: topic evolution in nanoscience[J]. Scientometrics, 2018, 115(1): 893－911.
[27] HUANG M H, YANG H W, CHEN D Z. Increasing science and technology linkage in fuel cells: a cross citation analysis of papers and patents[J]. Journal of informetrics, 2015, 9(2): 237－249.
[28] 吴菲菲，黄鲁成，石媛嫄. 基于文献和专利相互引用的科学与技术关系分析[J]. 科学学与科学技术管理，2013，34(10): 13－20.
[29] 彭彦淇，覃佳慧，叶鹰. 石墨烯研究中专利与论文的交叉引用分析[J]. 情报理论与实践，2018，41(7):18－21.
[30] 黄鲁成，王静静，李欣，等. 基于论文和专利的钙钛矿太阳能电池的技术机会分析[J]. 情报学报，2016，35(7):686－695.
[31] 陈二静，姜恩波. 文本相似度计算方法研究综述[J]. 数据分析与知识发现，2017，6(6): 1－11.

作者贡献说明：

俞琰:提出研究思路,设计研究方案,进行试验,撰写论文;
陈磊:数据清洗;
姜金德:分析数据,修改论文;
赵乃瑄:修改论文。

Patent Term Extraction by Integrating Keyword Knowledge From Paper

Yu Yan[1,2] Chen Lei[1] Jiang Jinde[3] Zhao Naixuan[1]
[1] Information Service Department, Nanjing Tech University, Nanjing 210009
[2] Computer Science Department, Southeast University Chengxian College, Nanjing 211816
[3] School of Business, Nanjing Xiaozhuang University, Nanjing 211171

Abstract: [Purpose/significance] In order to make up for the shortcomings of the patent text collection itself to limit the effect of patent term extraction, this paper proposes to use the rich keyword knowledge to obtain effective features outside the patent text to improve the patent term extraction effect. [Method/process] According to the keyword knowledge of related papers, two kinds of characteristic, degree of domain relevance and degree of head & tail are proposed to measure the possibility that candidate terms become terminology, and these characteristics are incorporated into the traditional method of patent term extraction. [Result/conclusion] The experimental results show that the degree of domain relevance and the degree of head & tail of the candidate terms obtained by using the keyword information of the papers make the method of combining the keyword knowledge of the papers significantly higher than the accuracy of the traditional term extraction method.

Keywords: patent term extraction paper keyword